Appendix 1 Expenditure-based Small Area Estimation

The poverty mapping analysis undertaken was based upon a statistical technique, sometimes referred to as small area estimation. This combines household welfare survey and Census data (both collected at approximately the same time) to estimate welfare or other indicators for disaggregated geographic units such as communities. Researchers at the World Bank initiated this approach in 1996 (Hentschel and Lanjouw, 1996). Refining the techniques continues with many collaborators. There is now considerable reference material, some available on the Internet, for readers interested in the details of this methodology (e.g. Hentschel et al. 1998, Hentschel et al. 2000, Statistics SA 2000, Alderman et al. 2002, Elbers et al. 2002, Elbers et al. 2003a and 2003b, Demombynes et al., 2002, Demombynes et al., 2003 and Mistiaen et al. 2002). Here, we give a relatively brief and non-technical summary of the approach[1].

The approach begins with the national representative household welfare survey to acquire a reliable estimate of household expenditure (y). This enables calculation of more specific poverty measures linked to a poverty line. Log-linear regressions model per capita expenditure using a set of explanatory variables (x) that are common to both the integrated household survey and the Census (e.g. household size, education, housing and infrastructure characteristics and demographic variables). These first-stage regression models are represented at the lowest geographical level for which the integrated household survey data is representative (Region), and a different first-stage model is estimated for each stratum (e.g. Region, urban, rural). Next, the estimated coefficients from these regressions (including the estimated error terms associated with those coefficients) are used to predict log per capita expenditure for every household in the Census. The household-unit data is then aggregated to small statistical areas, such as Counties, to obtain more robust estimates of the percentage of households living below the poverty line. These poverty rates may produce a poverty map showing the spatial distribution of poverty at the County level, in the case of Uganda, which represents a significantly higher level of resolution than the Region-level measures obtainable from using the integrated household survey alone.

In the first Uganda stage, variables within the Census and welfare monitoring surveys were examined in detail. The objective of this stage was to determine whether the variables were statistically similarly distributed over households in the population Census and in the household sample survey. For example, there are questions in both the population Census and in the HIS survey about household size, level of education of the household head, and type of housing. However, the exact questions and manner in which the answers are recorded differ in some cases e.g. the exact number of years of schooling for the household head was asked and recorded in the survey, while whether they have an education at a primary, secondary, or higher level is what was recorded in the Census. In many cases, there were also discrepancies between identically defined variables due to Regional variation in interpretation, rendering certain variables comparable in some Regions and not in others.

The next step was to investigate whether these common variables were statistically similarly distributed over households in the population and those sampled by the survey. This assessment was based on the following statistics for each variable obtained from both the survey and the Census for each stratum: (i) the mean, (ii) the standard error, (iii) and the values for the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles. First, the Census mean for a particular variable was tested to see if it lay within the 95 percent confidence interval around the household survey mean for the same variable. Second, for dummy variables, means were checked to ensure they were not smaller than three percent and not larger than 97 percent, so that the variables constructed contain some variation across households. Okwi et al., 2003 shows the results of the comparison of variable means for the Census and survey, by Region and for Urban and Rural areas. In general, there are between 23 and 33 variables sufficiently comparable to be included in the analysis.

The modelling steps of the analysis involved developing eight models, four rural and four urban (representing the four Regions), using the integrated household survey data in a regression analysis. The variable we were trying to explain in each model was per capita household expenditure for a household in a particular location. The independent or explanatory variables for the model were those observable household characteristics found as comparable variables in both the survey and the Census, as described above.

Combing the estimated first stage parameters with the observable characteristics of each household in the Census generated predicted per capita household expenditures (including an error estimate) for every household in the Census. For each model estimated, a stepwise regression procedure in SAS was used to select the subset of variables from the set of "comparable" variables that provided the best explanatory power for log per capita expenditure. A significance level criterion was chosen with no ceiling on the number of variables selected. All household survey variables that were significant at the five percent level were selected for the regression. The results of the regression analysis show that the models were quite successful at explaining the variation in household expenditures in both urban and rural areas. The adjusted R^2 ranged from .56 to .63 in urban areas, and from .31 to .44 in rural areas (with location means included). Despite not being very high, particularly in the rural areas, the explanatory power of the models is comparable to those attained elsewhere in Africa[2].

In general, in our specification, the following variables: household size, level of education, age of head of household, housing characteristics and district dummies plus interaction terms with certain household level variables, turned out to be key variables chosen in most regressions. As expected, household size had a negative correlation with household per capita expenditure. The housing variables showed mixed results depending on the strata. However, since these regressions are association models, the parameter estimates of the dependent variables cannot be interpreted as causal effects, but simply provide information on the direction of relationship.

From the first stage results, the relatively low R^2s in the rural areas may be attributed to at least two reasons. First, the number of variables in the Census' short forms is limited to mostly household composition, education and ethnic origin[3]. Though this information is correlated to e.g. family labour or ability to understand extension information other variables of obvious importance to rural households are not available such as: plot size, presence of livestock, soil quality or access to markets. Second, household composition and education only change slowly over time. The returns to agriculture are variables much dependent on rainfall, illness of family labourers, incidence of pests and diseases and prices. Again some of this variation may be captured, for instance the age of the head of household and proneness to disease are correlated, but much of the cross sectional variation attributable to any of these sources will remain unexplained and gets subsumed in the error term.

1 This section comes from CBS, 2003, with permission from the authors.

2 In comparison, the adjusted R2 ranges from 0.32 to 0.49 in urban areas and from 0.31 to 0.49 in rural areas of Kenya (CBS, 2003), from 0.27 to 0.55 in Mozambique, 0.45 to 0.77 in Ecuador, and from 0.445 to 0.638 in urban areas and 0.239 to 0.460 in rural areas in Madagascar (Mistiaen et al., 2002).

3 Inclusion of all the variables from the short form raised the R2 but not to the urban strata levels implying we still needed to use more information such as access to roads and markets to improve them.